Fundamentals of Working with Data

monia.ranalli@uniroma2.it

Exercise 1

Consider the following set of observations:

0, 43, 70, 14, 22, 13, 90, 50

1. Find the mean.
2. Find the median.

Exercise 2

The owner of a company in downtown Rome is concerned about the large use of gasoline by her employees due to urban sprawl, traffic congestion, and the use of energy inefficient vehicles such as SUVs. She'd like to promote the use of public transportation. She decides to investigate how many kilometers her employees travel on public transportation during a typical day. The values for her 10 employees (recorded to the closest kilometer) are

0,0,4,0,0,0,10,0,6,0

1. Find and interpret the mean, median and mode.
2. She has just hired an additional employee. He lives in a different city and travels 90 kilometers a day on public transport. Recompute the mean and median. Describe the effect of this new observation.

Exercise 3

The table summarizes responses of 4383 subjects in a recent General Social Survey to the question, *"Whitin the past months, how many people have you known personally that were victims of homicide?"*

Number of Victims	Frequency
0	3944
1	279
2	97
3	40
4 or more	23
Total	**4383**

1. To find the mean, it is necessary to give a score to the “4 or more” class. Find it, using the score 4.5.
2. Find the median. Is the “4 or more” class problematic for it?
3. If 1744 observations shift from 0 to 4 or more, how do the mean and median change?
4. Why is the median the same for parts 2 and 3, even though the data are so different?

Exercise 4

Consider the following two sets of observations:

Set 1: 2,3,3,3,4,4,4
Set 2: 2,3,3,3,3,3,4

1. Find the variance for each data set.
2. Which data set shows more variability?

Exercise 5

A company decides to investigate the amount of sick leave taken by its employees. A sample of 8 employees yields the following numbers of days of sick leave taken in the past year

0,0,4,0,0,0,6,0

1. Find and interpret the range.
2. Find and interpret the standard deviation s.
3. Suppose that 6 was incorrectly recorded and is supposed to be 60. Redo parts 1 and 2 with the correct data and descibe the effect of this outlier.

Exercise 6

The mean and standard deviation of a sample may change if data are rescaled.

1. Scores on a difficult exam have a mean of 57 and a standard deviation of 20. The teacher boosts all the scores by 20 points before awarding grades. Report the mean and the variance of the boosted scores.
2. Referring to point 1, what happens to the mean if the students get a grade rise of 3%?
3. Suppose that the annual income for some group has a mean of \$ 39,000 and a standard deviation of \$ 15,000. Values are converted to euros. If one euro equals \$2.00, report the mean and standard deviation in European currency.

Exercise 7

A professor examined the results of the first exam given in her statistics class. The scores were

70, 84, 59, 73, 86, 35, 81, 75.

1. Find the mean and the median.
2. Would you guess that the distribution is skewed or roughly symmetric? Why?
3. Find the standard deviation.

Exercise 8

For the question "How many children have you ever had?", the results were

No.Children	0	1	2	3	4
Count	25	15	20	5	0

1. Find the variance and the standard deviation.

Exercise 9

The data values below represent the prices per share of the 20 most actively traded stocks on the New York Stock Exchange (rounded to the nearest dollar) on February 18 2011.

5,15,2,16,5,5,21,33,19,9,7,9,48,39,52,17,85,13,35,10

1. Construct a histogram.
2. Find the median, the first quartile, and the third quartile.
3. Sketch a box plot. What feature of the distribution displayed in the histogram is not obvious in the box plot? (Hint: Are there any gaps in the data?)

Exercise 10

The fertility rate for a nation is measured as the average number of children per adult woman. The table below shows results for western European nations, the United States, Canada, and Mexico, as reported by the United Nations in 2005.

Country	Fertility	Country	Fertility
Austria	1.4	Netherlands	1.7
Belgium	1.7	Norway	1.8
Denmark	1.8	Spain	1.3
Finland	1.7	Sweden	1.6
France	1.9	Switzerland	1.4
Germany	1.3	United Kingdom	1.7
Greece	1.3	United States	2.0
Ireland	1.9	Canada	1.5
Italy	1.3	Mexico	2.4

1. Find the quartiles (Q_1, Q_2, Q_3) for the fertility rates.

2. Find the interquantile range (IQR).

3. Find the five-number summary.

Exercise 11

The 2007 unemployment rates of countries in the European Union are shown in the table below.

Country	Unemployment rate	Country	Unemployment rate	Country	Unemployment rate
Belgium	7.8	France	8.4	Italy	6.7
Denmark	3.2	Portugal	7.2	Finland	7.0
Germany	7.7	Netherlands	3.6	Austria	4.5
Greece	8.7	Luxembourg	5.0	Sweden	6.0
Spain	8.6	Ireland	4.4	U.K.	5.4

1. Identify the five-number summary, and sketch a box plot.

2. Provide a graphical representation of the distribution.

3. Greece had the highest unemployment rate of 8.7. Is it an outlier? Explain.

4. Find the mean and standard deviation.

5. What unemployment value for a country would have a z-score equal to 0?

Exercise 12

Looking at the following table

Gender	Binge Drinker	Non-Binge Drinker	Total
Male	1908	2017	**3925**
Female	2854	4125	**6979**
Total	**4762**	**6142**	**10904**

1. Identify the response variable and the explanatory variable.

2. Report the cell counts of subjects who were (i) male and a binge drinker, (ii) female and a non-binge drinker.

3. Construct a contingency table that shows the conditional proportions of sampled subjects who do or do not binge drink, given gender.

4. (*challenging question*) Based on part 3, does it seem that there is an association between binge drinking and gender?

Exercise 13

The table shows results of whether the death penalty was imposed in murder trials in Florida between 1976 and 1987. For instance, the death penalty was given in 53 out 467 cases in which a white defendant had a white victim.

Victim's Race	Defendant's Race	YES Death Penalty	NO Death Penalty	Total
White	White	53	414	**467**
White	Black	11	37	**48**
Black	White	0	16	**16**
Black	Black	4	139	**143**

1. Consider only the cases in which the victim was white. Find the conditional proportions that got the death penalty when the defendant was white and when the defendant was black. Describe the association.

2. Repeat part 1 for cases in which the victim was black.

3. Construct a summary contingency table that describes the association between the death penalty verdict and defendant's race, ignoring the information about the victim's race.

4. (*challenging question*) Find the conditional proportions and describe the association.

Exercise 14

Consider the data

x	y
3	8
4	13
5	12
6	14
7	16

1. Sketch a scatterplot.

2. Would you expect a positive association, a negative association or no association between x and y?

3. Compute the correlation coefficient, r.

Exercise 15

An instructor of Statistics collected data from one of her classes in Spring 2016 to investigate the relationship between Study time per week (number of hours) to predict the final grade. For the 8 students in her class the data were as shown in the table.

Student	**Study Time**	**Grade**
1	14	26
2	25	30
3	15	20
4	5	18
5	10	23
6	12	25
7	5	21
8	21	28

1. Identify the response variable and the explanatory variable.

2. Construct a scatterplot.

3. Find and interpret the correlation.